Capstone Project 1 - Dental Health in Youth

Client and Problem:

What is a good indicator for dental health across the globe? Which countries and areas have the largest amount of oral disease in children under 12 and what is most correlated? The goal of this project is to build models to estimate and predict tooth issues using many demographic factors for physicians and organizations who want to understand what kind of demographics and conditions that contribute to dental issues in children. This can then be used to properly allocate funding and focus efforts.

According to the World Health Organization, the most common oral diseases are dental cavities, periodontal (gum) disease, oral cancer, oral infectious diseases, trauma from injuries, and hereditary lesions. They also state that 60–90% of children and nearly 100% of adults have dental cavities across the globe. Can we group similar countries together based on demographic data and find a trend in the number of childhood dental problems?

Data: The data is acquired thru Gapminder and WHO website from the different origins -

Badteeth.csv

Bad teeth per child under 12 years old, average by country.
Source - WHO

Gdp.csv

GDP per capita by country in US dollars and inflation adjusted.
Source - World Bank

Healthexpend.csv

Government health spending per person in US dollars by country.
Source - WHO

Sugar_consumption.csv

Sugar consumption per person in grams per day by country.
Source - FAO

Literacy.csv

Percent of adults age 15 and up who are literate by country.
Source - UNESCO

Water.csv

Percent of population with access to basic drinking water sanitation by country.
Source - WHO

Tobacco.csv

Percentage tobacco use by youth, aged 17 or younger by country.

Source - WHO

Smokers.csv

Percent smoking rate by youth, aged 17 or younger by country.
Source - WHO

Lowbmi.csv

Percent of children more than 2 std below the median BMI by country.
Source - WHO

Adbirthrate.csv

Adolescent birth rate per 1000 women age 15-19 by country.
Source - WHO

Tools Used:

Pandas - read and save files, convert types,

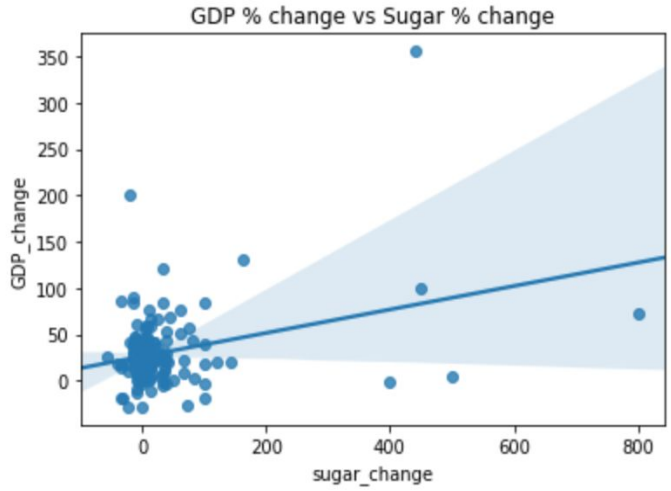
Seaborn, matplotlib - visualizations

Sklearn - Machine Learning, linear regression, clustering, PCA

Scipy - correlations, significance tests, dendrogram, linkage, fcluster

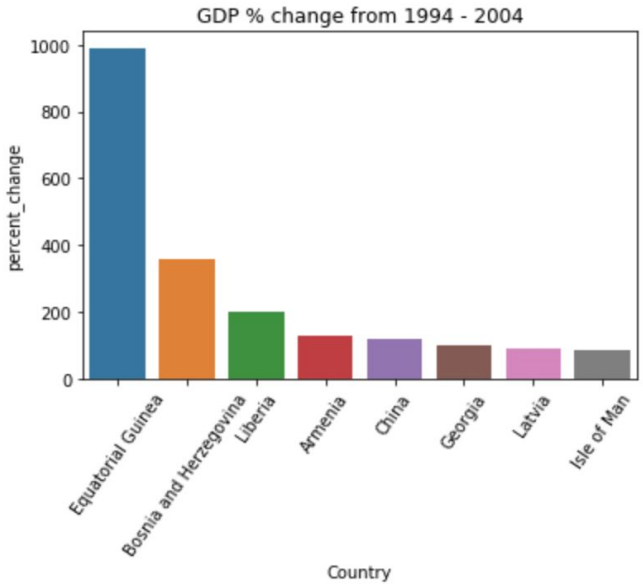
Exploring the Data:

It makes sense that wealthier countries would have higher health across the board, but wealthier countries also consumed more sugar. I wanted to explore the changes between GDP and sugar consumption over time.

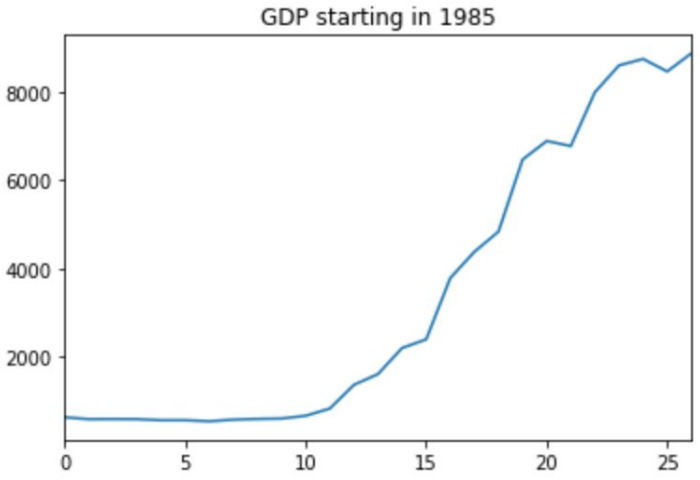


Between 1994 and 2004, the percentage change in GDP compared to change in sugar consumption is shown below. While there is a correlation, most countries seemed to stay within a 100% increase or less over the ten years.

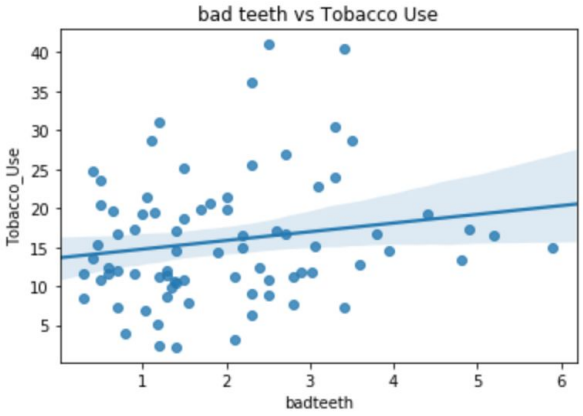
While most countries' GDP increased, some increased at a much higher rate. The largest changes in GDP between 1994 and 2004 are shown. Equatorial Guinea had a 991% change in GDP.



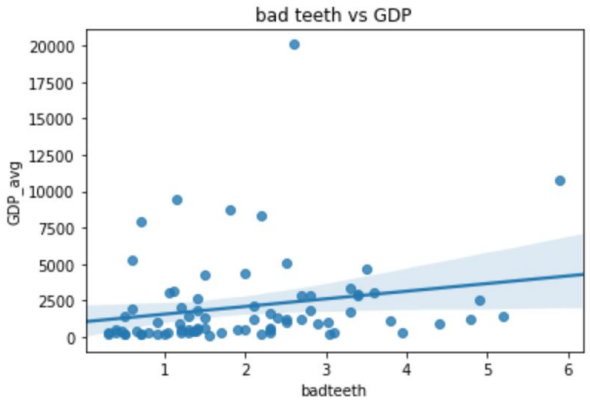
Because Equatorial Guinea showed such a startling difference, I wanted to make sure it was not a typo or mistake in the data by actually showing the change in GDP over the years. This country seemed to have a very constant GDP until around 1995 when it showed a very steady increase into the early 2000's. Judging by this, it is safe to conclude that Equatorial Guinea has simply raised their GDP significantly and is not mistake in the data. This spike in GDP is likely contributed by the discovery of large oil preserves in the country in the 90's. Equatorial Guinea continues to have the highest per capita income in Africa.

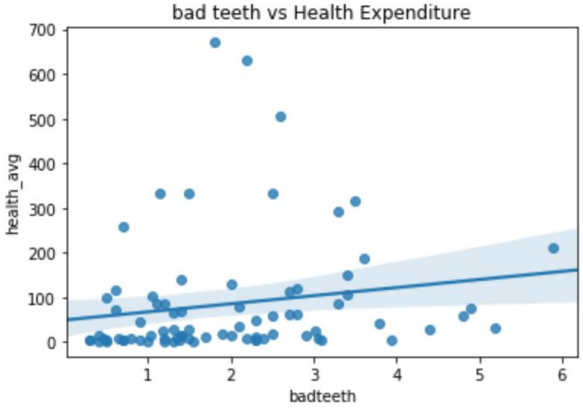


One variable that I expected to have more of an effect on dental health was tobacco. There is barely a correlation, but this is most likely due to the dental records being from children who, even if they did use tobacco, would not have long enough exposure to affect teeth.



While tobacco use affects people later in life, GDP and health expenditure has the potential to affect you even before birth.





Both GDP and health expenditure has a slight positive correlation on the amount of dental issues. The three countries studied with the highest average GDP between 1994 and 2004 are Cyprus, Saudi Arabia and Kuwait.

Country	badteeth	literacy_avg	water_avg	GDP_avg	sugar_avg	health_avg	Low_BMI	Adolescent birth rate	Tobacco_Use	Smokers
Cyprus	1.14	95.583051	100.0	9424.547896	90.161136	334.533933	0.9	4.2	19.5	13.9
Saudi Arabia	5.90	75.086297	97.0	10724.992842	63.511591	212.081834	7.3	17.6	14.9	8.9
Kuwait	2.60	69.992105	100.0	20102.283588	103.362727	506.458186	3.1	7.1	17.0	15.9

These countries have widely varying literacy rates and sugar consumption rates, but high GDP and health expenditures and fairly average other factors. Cyprus' number of bad teeth falls near the 25% range, Kuwait's number of bad teeth is almost to the 75% range, and Saudi Arabia is the highest on the list.

Data Wrangling:

There are 10 different files that all needed to be cleaned and prepared. Half of them came from a Kaggle dataset and were very similar. These files had information for a large number of countries, each with records over a span of 2 and 20 years, littered with missing data and odd formatting. The other half came from the WHO website. Similar issues needed to be fixed within these files such as dropping missing values, converting dtypes, changing indexes, and then averaged over time. All data after 2004 has been discarded (as the dental data being compared is from 2004). The separate files were then merged.

The first 5 rows in the merged dataframe are as follows:

Country	badteeth	literacy_avg	water_avg	GDP_avg	sugar_avg	health_avg	Low_BMI	Adolescent birth rate	Tobacco_Use	Smokers
Albania	3.02	98.712978	88.0	1011.752695	49.504318	24.751387	1.9	19.7	11.8	11.5
Algeria	2.30	59.752193	90.4	1625.140989	67.995227	50.424490	6.2	12.4	9.0	5.7
Angola	1.70	67.405416	38.4	318.739949	33.129091	12.275108	10.3	190.9	19.8	2.3
Antigua and Barbuda	0.70	98.950000	98.0	7874.822798	98.443182	259.996606	3.5	66.8	11.9	7.4
Belarus	2.70	98.737052	98.0	1196.023043	100.526923	64.079201	2.5	21.6	26.9	26.5

Then Pearson correlations in the resulting dataframe are found.

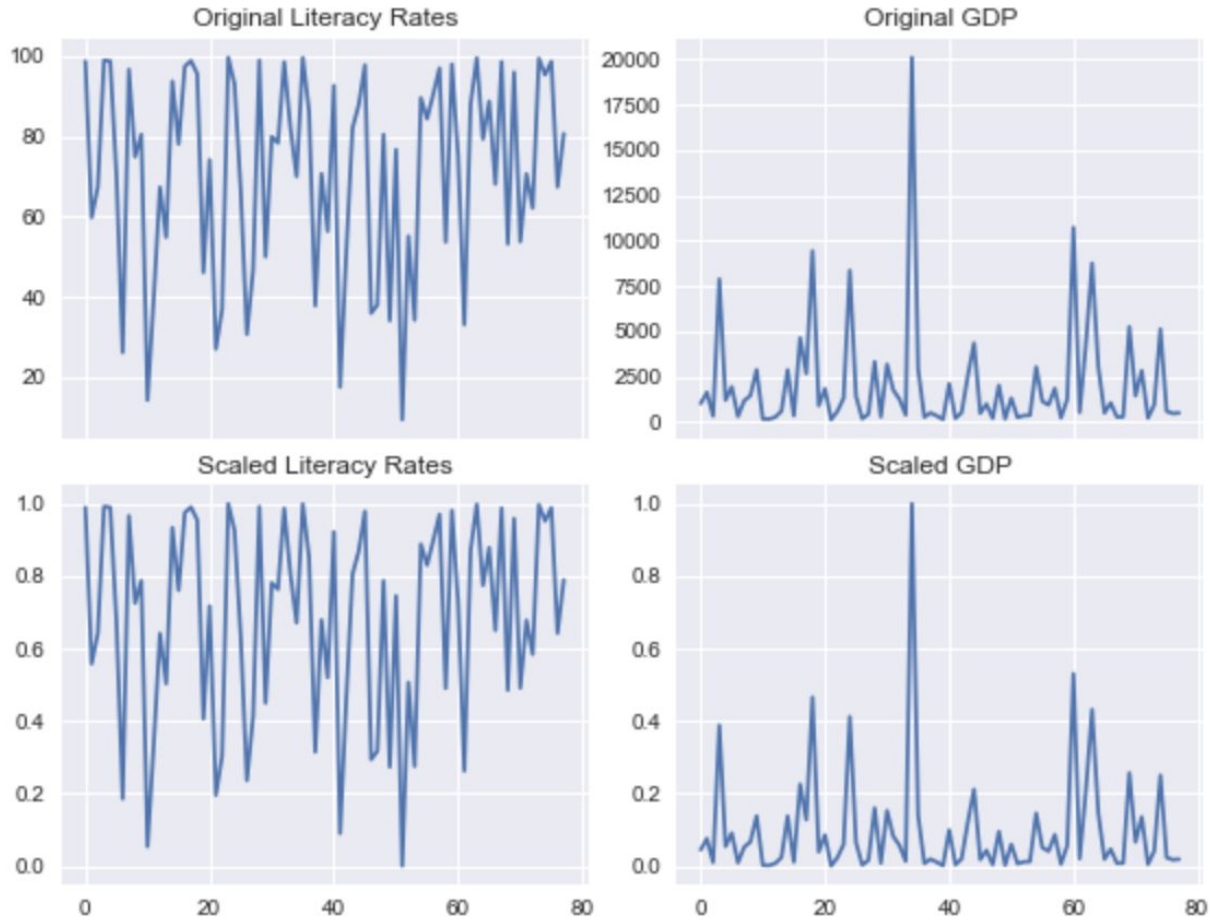
	badteeth	literacy_avg	water_avg	GDP_avg	sugar_avg	health_avg	Low_BMI	Adolescent birth rate	Tobacco_Use	Smokers
badteeth	1.000000	0.325372	0.401619	0.213518	0.320137	0.167307	-0.251279	-0.299398	0.173012	0.363074
literacy_avg	0.325372	1.000000	0.636242	0.340530	0.557734	0.417457	-0.527774	-0.567905	0.170547	0.440038
water_avg	0.401619	0.636242	1.000000	0.514874	0.702112	0.528664	-0.383963	-0.752723	0.175181	0.479703
GDP_avg	0.213518	0.340530	0.514874	1.000000	0.430608	0.830989	-0.332287	-0.411018	0.118022	0.328629
sugar_avg	0.320137	0.557734	0.702112	0.430608	1.000000	0.404256	-0.375937	-0.472294	0.287263	0.527458
health_avg	0.167307	0.417457	0.528664	0.830989	0.404256	1.000000	-0.384411	-0.415634	0.173532	0.421346
Low_BMI	-0.251279	-0.527774	-0.383963	-0.332287	-0.375937	-0.384411	1.000000	0.253361	-0.101926	-0.434536
Adolescent birth rate	-0.299398	-0.567905	-0.752723	-0.411018	-0.472294	-0.415634	0.253361	1.000000	-0.042804	-0.331288
Tobacco_Use	0.173012	0.170547	0.175181	0.118022	0.287263	0.173532	-0.101926	-0.042804	1.000000	0.713721
Smokers	0.363074	0.440038	0.479703	0.328629	0.527458	0.421346	-0.434536	-0.331288	0.713721	1.000000

While many of our features are correlated, nothing seems to stand out as highly correlated to our number of bad teeth in youth.

Data Preprocessing

The next step was to normalize the data. Using Min-Max Scaling, the data is scaled to a fixed range between 0 to 1.

This creates smaller standard deviations and lessens the effects of outliers. Scaling the data is important because the model is sensitive to magnitude, and the units of each feature is different. Scaling makes the data more comparable. An example of how scaling changes the data is shown below with Our original literacy rates and GDP for each country above the scaled data.



As you can see, the distribution of the data has not changed, but now you are not comparing Albania's 98.7% literacy rate with a GDP of about $1011.8, you're using 0.989 for literacy rate and 0.044 for GDP. This automatically gives you a sense that Albania has a relatively high literacy rate and a low GDP compared to the other countries.

To read more about min/max scaling,
http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Approach

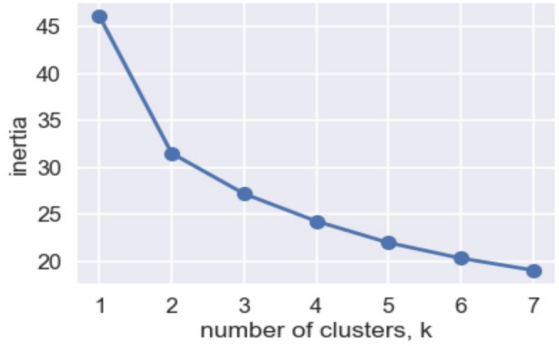
We do not have ground truths or expert analysis, only surveys. Because of this I chose to use an unsupervised learning method - Clustering.

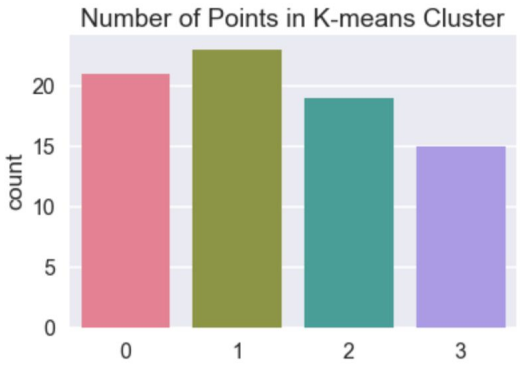
Clustering
K-Means

Now that we have properly scaled the data, we can cluster them. The first method I used is K-means clustering. K-Means clustering works by randomly selecting k number of points in your data to be your temporary centers. The clusters are then created by assigning each data point

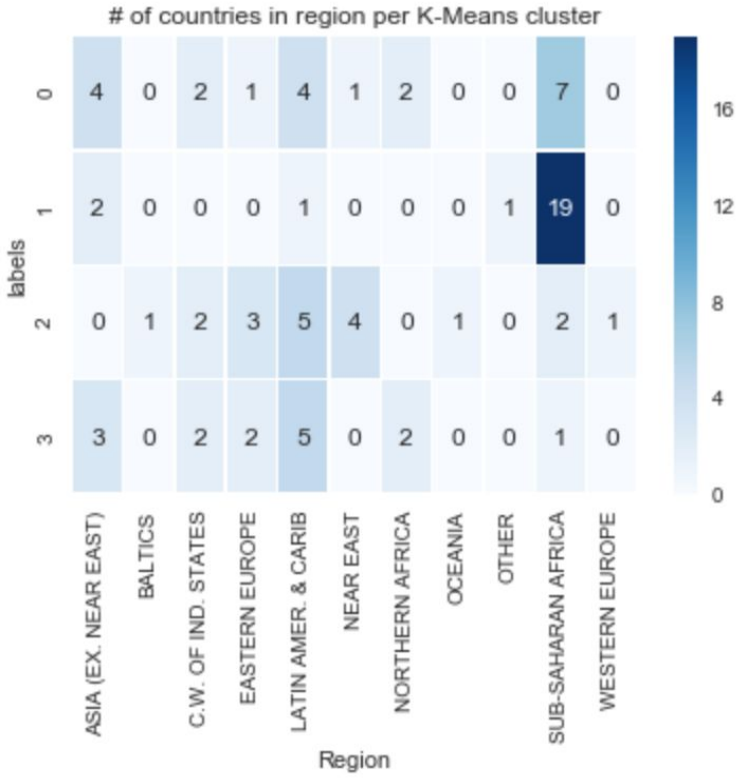
to the nearest center. The algorithm evaluates a new center of this created cluster and the process is repeated until it reaches convergence.

There are many different ways to choose the number of clusters. I chose to use the “elbow method” which works by measuring the inertia (a measure of variation) of each number of clusters. Most cases this creates a curve that shows when adding more clusters no longer helps reduce the inertia effectively. I chose to use 4 clusters, which seems to have similar size clusters.



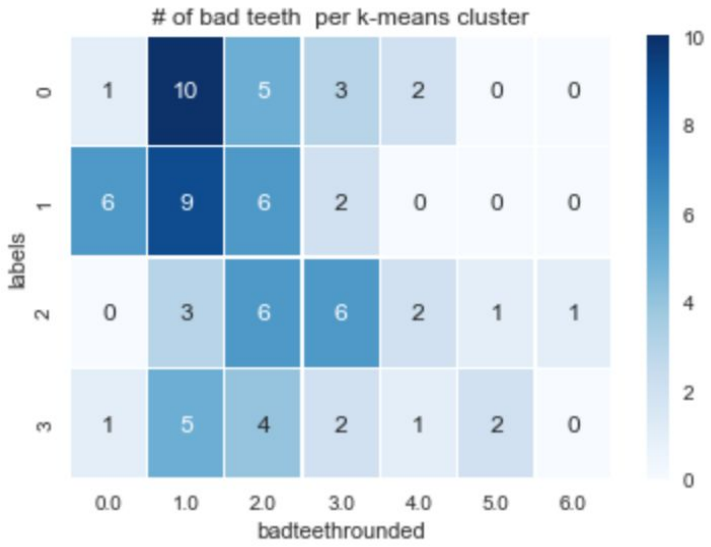


With a list of cluster labels from our model, I created a table showing how many countries from each region was assigned to each cluster.

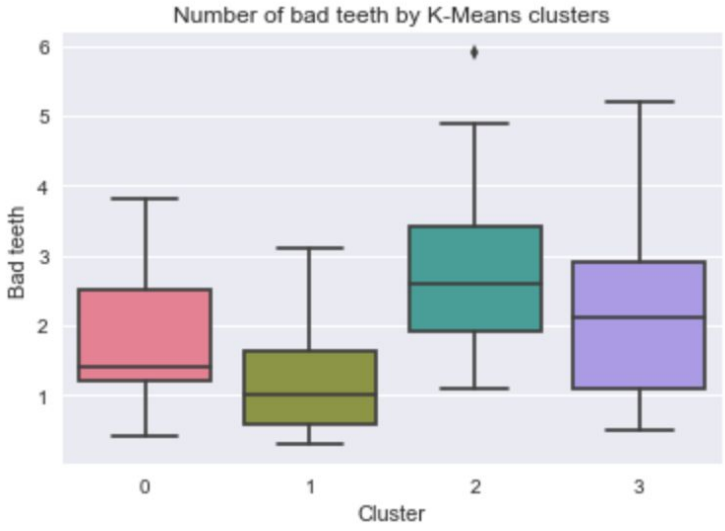


It looks like cluster 1 is almost entirely made up of sub-saharan countries and the northern African countries are split between cluster 0 and 3. Other than this there does not seem to be any strong links between cluster and region.

Next I wanted to compare the number of bad teeth in each K-Means cluster.



Comparing the number of bad teeth in each cluster is a little bit more telling than the region was. It looks like cluster 1 (which was made up of mostly sub-saharan Africa) generally has the least, then cluster 0 has very little (but more than 1), and cluster 2 has the higher values. Cluster 3 seems to span the middle range.



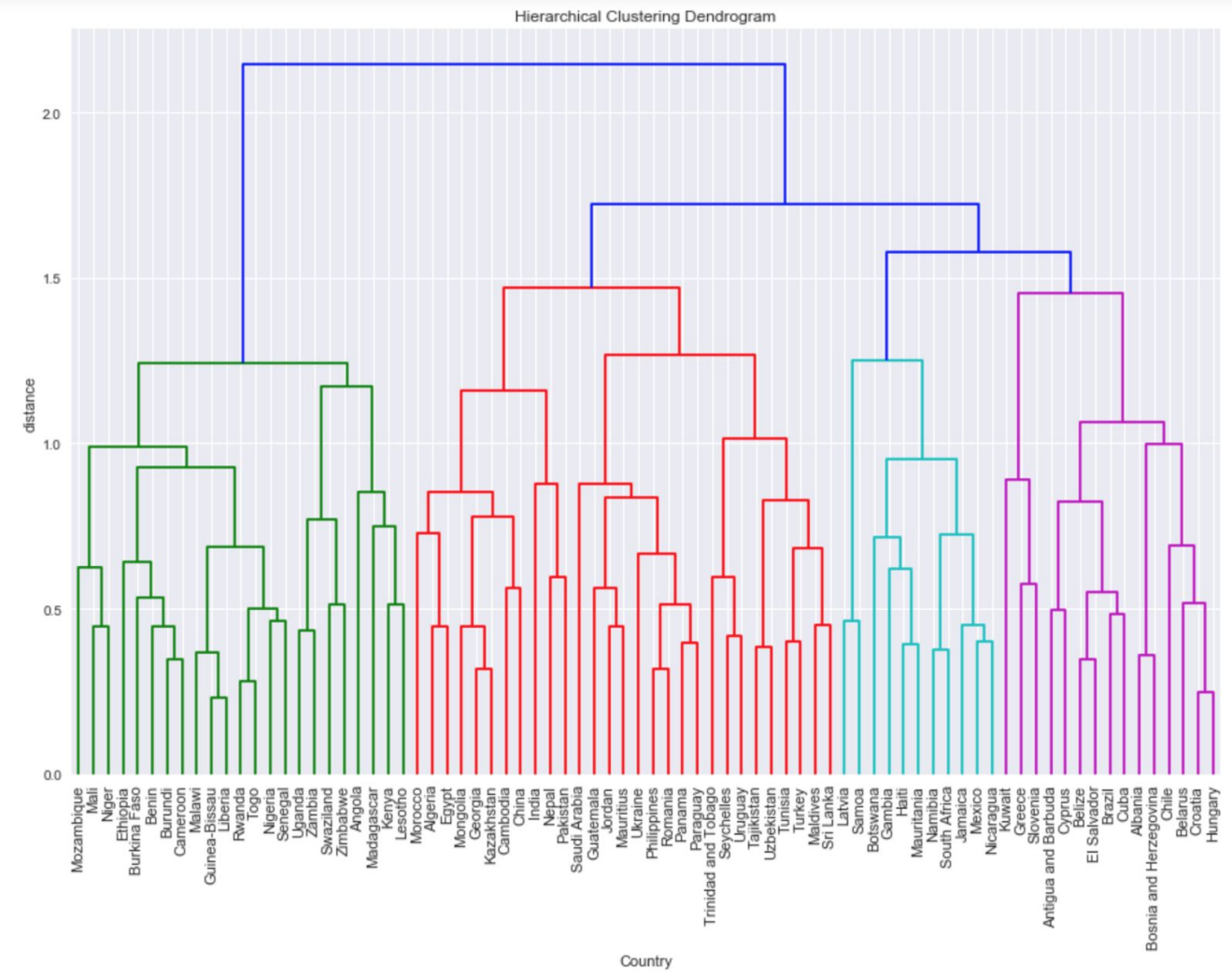
Box plots are a better for comparing means and distributions. This shows that there are definitely those trends found earlier and also shows an outlier. This particular outlier is Saudi Arabia and would be an outlier regardless of which cluster it was assigned. The median values are:

- Cluster 0 - 1.4
- Cluster 1 - 1.0
- Cluster 2 - 2.6
- Cluster 3 - 2.1

You can find more on K-Means Clustering at
https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html

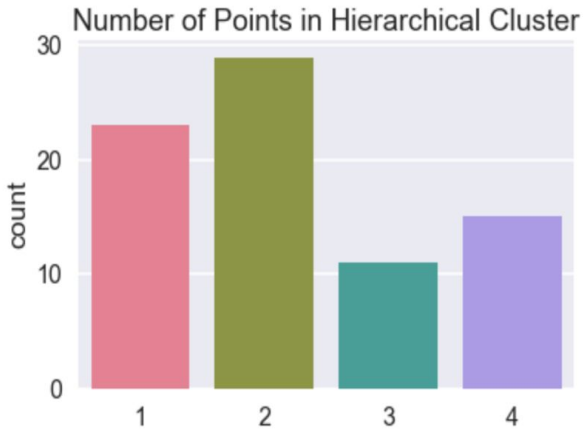
Hierarchical Clustering

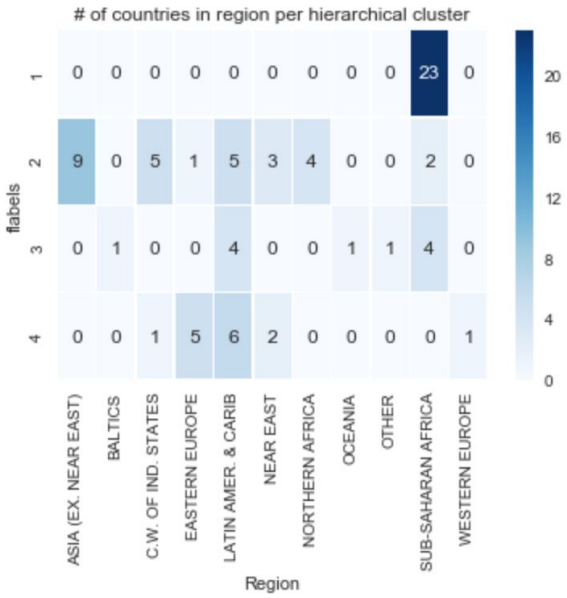
I decided to try a different clustering method. The complete linkage hierarchical clustering works by starting with each element as its own cluster. In this case, each Country starts as a cluster, then the closest clusters get combined. These new clusters then are merged with the closest clusters and so on. What makes it a “complete” linkage is that the distances between the clusters are measured by the furthest two points in the clusters. You decide how many clusters to choose by changing the distance at which you stop at. A dendrogram is a great way to visualize this.



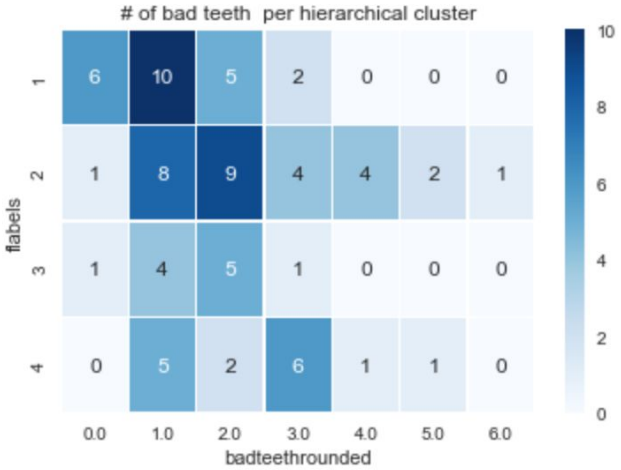
In order to create 4 clusters, like the K-means clustering method, the distance to stop at was 1.5. If we had chosen a distance of 2 instead, cluster 1 would remain the same and the other three would be combined into one cluster.

Now this way of clustering gives very different results than our K-means method.





The Hierarchical clustering also clustered the majority of Sub-saharan Africa together. In fact all of the first cluster consists of is Sub-saharan African countries. The number of countries per cluster varies slightly more in this clustering than with K-Means, with the smaller cluster having only 11 countries.

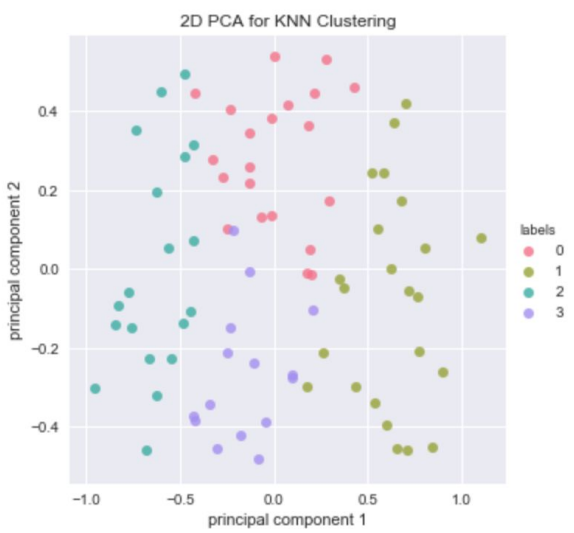


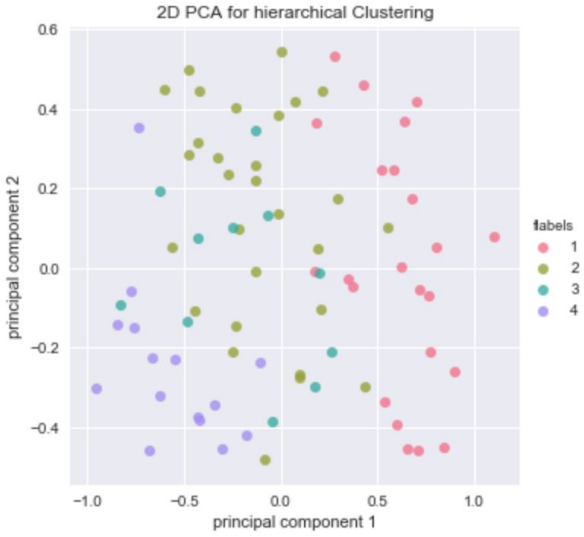
As far as the number of bad teeth per cluster, there are many similarities to the K-Means method.

More information on Hierarchical clustering can be found at
https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/

PCA

PCA, or Principal Component Analysis is used in this case to reduce the number of dimensions from 10 variables to 2 so we can plot the clusters on a 2d graph. I chose not to use the PCA information for the clustering because I did not have too large of data, and after finding the cumulative explained variance ratio I found that reducing to 2 components would retain only 57.7% of the variance from the original data.





For more on PCAs:
https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

Trends

Given the low amount of data, there was some struggle with finding useful results. In general, wealthier countries with more access to clean water and higher literacy rates and overall health had experienced slightly more teeth issues in children.

Sub-saharan Africa was very distinct from the other regions and the countries were quite similar to one another. The seven countries with the least bad teeth are all from Sub-saharan Africa. This is not across the board though, as Mauritius has the third highest child dental issues.

The Hierarchical clustering seemed to cluster the extremes of the bad teeth very well. The ten countries with the largest amount of childhood tooth problems were all grouped into 2 clusters and the lowest six countries were all in the same cluster.

The World Health Organization website claims that risk factors for oral diseases include unhealthy diet, tobacco use and alcohol use. All of these are easier to access with a wealthier population. I am going to suggest, given the data, that social determinants and general culture

around dental hygiene explains a large portion of the variance of oral health in children. Countries with 5 times the amount of health expenditure per person have the same if not more oral issues than other countries, and no variable has a remarkable correlation. This would also explain why some regions are so different from others.

Recommendations for Further Analysis - Further analysis would benefit from more data, which could come in the form of more countries included, data from more recent years, and/or more variables to compare. I would also love to see how much different dental procedures cost in each region to compare a return of investment or compare the availability of current oral health services. Additionally, the WHO claims that fluoride in drinking water can significantly reduce cavities in the entire population. I would recommend doing more research or conducting a survey to see which countries currently have flouride in their public drinking water and how much of the population has access to it. I would love to see if there is actually a correlation in the data. If possible it might even be beneficial to break down the country further into sections as to reduce the misrepresentation of a country with a large wealth gap.